NANO-MATERIAL CONFIGURATION DESIGN WITH DEEP SURROGATE LANGEVIN DYNAMICS

Thanh V. Nguyen
Iowa State University
thanhng@iastate.edu

Youssef Mroueh, Samuel Hoffman, Payel Das & Pierre Dognin
IBM Research
{mroueh@us, shoffman@, daspa@us, pdognin@us}.ibm.com

Giuseppe Romano
MIT
romano@mit.edu

Chinmay Hegde
New York University
chinmay.h@nyu.edu

ABSTRACT

We consider the problem of optimizing by sampling under multiple black-box constraints in nano-material design. We leverage the posterior regularization framework and show that the constraint satisfaction problem can be formulated as sampling from a Gibbs distribution. The main challenges come from the black-box nature of the constraints obtained by solving complex and expensive PDEs. To circumvent these issues, we introduce Surrogate-based Constrained Langevin dynamics for black-box sampling. We devise two approaches for learning surrogate gradients of the black-box functions: first, by using zero-order gradients approximations; and second, by approximating the Langevin gradients with deep neural networks. We prove the convergence of both approaches when the target distribution is log-concave and smooth. We also show the effectiveness of our approaches over Bayesian optimization in designing optimal nano-porous material configurations that achieve low thermal conductivity and reasonable mechanical stability.

1 INTRODUCTION

In many real-world design problems, optimal designs simultaneously satisfy multiple conflicting constraints that can be expensive to evaluate. As an example, in computational material design the main goal is to fabricate new material configurations that meet a series of physical constraints. These constraints are often specified by Partial Differential Equations (PDEs) via black-box numerical solvers. Such solvers are complex, expensive to evaluate, and often offer no access to the inner variables or the gradients.

Black-box optimization. The black-box nature of the above problems prevents the use of gradient-based optimization, and several alternative approaches have been proposed. The first common approach is based on finite differences using Gaussian smoothing (or zero-order optimization) to estimate gradients (Nesterov & Spokoiny, 2017; Duchi et al., 2015; Ghadimi & Lan, 2013). An alternative for optimizing expensive black-box functions is Bayesian Optimization (BO) (Mockus, 1994; Jones et al., 1998; Frazier, 2018).

Black-box sampling. Similar to black-box optimization, the problem of sampling from a distribution with unknown likelihood that can only be point-wise evaluated is called black-box sampling (Chen & Schmeiser, 1998; Neal, 2003). Naturally, zero-order methods via Gaussian smoothing (Nesterov & Spokoiny, 2017) can be extended to black-box sampling, for example using Langevin dynamics (Shen et al., 2019). Compared with optimization, sampling approaches are natural for optimal design since one might prefer a distribution of candidate designs over one single point. However, it is expensive to repeatedly query the PDE solvers.

In this paper, we consider the problem of optimizing multiple black-box objectives as sampling from a Gibbs distribution with compact support. We show that the sampling problem can be cast in the

framework of *constrained* Langevin dynamics and extend the traditional Langevin dynamics to the black-box case with constraints and compact support.

To alleviate the computational burden in zero-order methods, we propose *Surrogate Model-Based Langevin dynamics* that consists of two steps: (i) Learning (using training data) a deep surrogate of the *gradient* of the potential of the Gibbs distribution. (ii) Using the surrogate model in the constrained Langevin dynamics *in lieu of* the black-box potential. The surrogate enables more efficient sampling.

To summarize, we make the following contributions: We cast the problem of optimizing a black-box function under constraints as sampling from a Gibbs distribution. We introduce Constrained Zero-Order Langevin Monte Carlo, using projection or proximal methods, and prove the convergence to the target Gibbs distribution. We introduce Surrogate Langevin Monte Carlo by learning surrogate gradient of the potential of the Gibbs distribution using deep neural networks. Finally we show the effectiveness of our approach in nano-porous configurations design with improved thermoelectric efficiency and mechanical stability.

2 Constraint Satisfaction As Sampling

Our goal is to find a posterior distribution q of samples satisfying a series of equality and inequality constraints: $\psi_j(x) = y_k, j = 1 \dots C_e$, and $\phi_k(x) \leq b_k, k = 1 \dots C_i$ where $x \in \Omega$ and $\Omega \subset \mathbb{R}^d$ is a bounded domain. We assume a prior distribution p_0 whose analytical form is known. The main challenge is that ψ_j and ϕ_k are black box functions obtained from solving complex PDEs. We choose Lagrangian parameters $\lambda_j > 0$ and obtain the relaxation:

$$\min_{q, \int_\Omega q(x)=1} \mathscr{L}(q) \tag{1}$$

where $\mathscr{L}(q) = \mathrm{KL}(q, p_0) + \sum_{j=1}^{C_e} \lambda_j \mathbb{E}_{x\sim q}(\psi_j(x) - y_k)^2 + \sum_{k=1}^{C_i} \lambda_k \mathbb{E}_{x\sim q}(\phi_k(x) - b_k)_+$.

Following the posterior regularization framework of Ganchev et al. (2010); Hu et al. (2018), we obtain that the constraint satisfaction problem corresponds to sampling from a Gibbs distribution:

Lemma 1. *The solution to the distribution learning problem given in Eq. 1 is given by:*

$$\pi(x) = \frac{\exp(-U(x))}{Z} \mathbb{1}_{x\in\Omega}, \tag{2}$$

where $U(x) = -\log p_0(x) + \sum_{j=1}^{C_e} \lambda_j(\psi_j(x) - y_k)^2 + \sum_{k=1}^{C_i} \lambda_k(\phi_k(x) - b_k)_+$ *and* $Z = \int_{x\in\Omega} \exp\left(-U(x)\right) dx$.

3 Constrained Langevin Dynamics

Sampling from $\pi(x)$ in Eq. 2 can be done using Langevin dynamics. The constrained white-box setting was recently studied in (Dalalyan, 2017; Bubeck et al., 2015; Brosse et al., 2017; Durmus et al., 2019). We give a quick review:

Assumption A: Ω is a convex set such that $0 \in \Omega$, Ω contains a Euclidean ball of radius r, and Ω is contained in a Euclidean ball of radius R. (e.g., Ω might encode box constraints.) Let $R = \sup_{x,x'\in\Omega} ||x - x'|| < \infty$.

Assumption B: We assume that U is convex, β-smooth, and with bounded gradients:

$$\|\nabla_x U(x) - \nabla_y U(y)\| \leq \beta \|x - y\|, \quad \forall x, y \in \Omega.$$
$$\|\nabla U(x)\| \leq L, \quad \forall x \in \Omega \text{ (Boundedness).}$$

The Total Variation (TV) distance between measures μ, ν is $\mathrm{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$. The projection onto Ω, $P_\Omega(x)$ is defined as follows: for all $x \in \Omega$, $P_\Omega(x) = \arg\min_{z\in\Omega} \|x - z\|^2$.

Projected Langevin dynamics. Similar to projected gradient descent, Bubeck et al. (2015) introduced Projected Langevin Monte Carlo (PLMC) and proved its mixing propreties towards the stationary distribution π. PLMC is given by the following iteration for $k = 0 \dots K-1$ **(PLMC)**:

$$X_{k+1} = P_\Omega\left(X_k - \eta\nabla_x U(X_k) + \sqrt{2\lambda\eta}\xi_k\right), \tag{3}$$

for $k = 0 \dots K-1$, where $\xi_k \sim \mathcal{N}(0, I_d)$, η is the learning rate, and $\lambda > 0$ is a variance term.

Proximal Langevin dynamics. Similar to proximal methods in constrained optimization, Brosse et al. (2017) introduced Proximal LMC (ProxLMC) that uses the iteration for $k = 0 \dots K-1$ **(ProxLMC)**:

$$X_{k+1} = \left(1 - \frac{\eta}{\gamma}\right) X_k - \eta \nabla_x U(X_k) + \frac{\eta}{\gamma} P_\Omega(X_k) + \sqrt{2\lambda\eta}\xi_k \tag{4}$$

where η is the step size and γ is a regularization parameter.

We denote by μ_K^{PLMC} and μ_K^{ProxLMC} the distributions of X_K obtained by iterating Eq. 3 and Eq. 4 respectively. Under Assumptions **A** and **B**, both these distributions converge to the target Gibbs distribution π in the total variation distance. In particular, Bubeck et al. (2015) showed that for $\eta = \tilde{\Theta}(R^2/K)$, we obtain:

$$\text{TV}(\mu_K^{\text{PLMC}}, \pi) \leq \varepsilon \text{ for } K = \tilde{\Omega}(\varepsilon^{-12} d^{12}). \tag{5}$$

Likewise, Brosse et al. (2017) showed we can obtain the following for $0 < \eta \leq \gamma(1 + \beta^2\gamma^2)^{-1}$:

$$\text{TV}(\mu_K^{\text{ProxLMC}}, \pi) \leq \varepsilon \text{ for } K = \tilde{\Omega}(\varepsilon^{-6} d^{5}), \tag{6}$$

where the notation $\alpha_n = \tilde{\Omega}(\beta_n)$ means that there exists $c \in \mathbb{R}, C > 0$ such that $\alpha_n \geq C\beta_n \log^c(\beta_n)$.

4 Black-Box Constrained Langevin

We now introduce our variants of black-box constrained LMC. We explore two strategies for approximating the gradient of $U(x)$ by (i) adopting derivative-free optimization and (ii) learning a *surrogate deep model* that approximates the gradient of the potential. Let $G : \Omega \to \mathbb{R}^d$ be a surrogate gradient that approximates the true gradient $\nabla_x U$.

Assumption C. The surrogate gradient G satisfies $\mathbb{E}\left\|G(Y_k)\right\|^2 < \infty, \forall k = 0, 1, \dots$.

Surrogate projected Langevin dynamics. Given Y_0, the Surrogate Projected LMC (**S-PLMC**) replaces the potential gradient $\nabla_x U$ in Eq. 3 with the surrogate gradient G:

$$Y_{k+1} = P_\Omega\left(Y_k - \eta \mathbf{G}(Y_k) + \sqrt{2\lambda\eta}\xi_k\right), k = 0 \dots K-1 \tag{7}$$

Surrogate proximal Langevin dynamics. Similarly, the Surrogate Proximal LMC (**S-ProxLMC**) replaces the unknown potential gradient $\nabla_x U$ in Eq. 4 with the gradient surrogate G, for $k = 0 \dots K-1$:

$$Y_{k+1} = \left(1 - \frac{\eta}{\gamma}\right) Y_k - \eta \mathbf{G}(Y_k) + \frac{\eta}{\gamma} P_\Omega(Y_k) + \sqrt{2\lambda\eta}\xi_k. \tag{8}$$

We now present our main theorems. We bound the total variation distance between the trajectories of the surrogate Langevin dynamics (S-PLMC, and S-ProxLMC) and the true LMC dynamics (PLMC and ProxLMC).

Theorem 1 (S-PLMC and S-ProxLMC Mixing Properties)**.** *Under Assumption* **C***, we have:*

1. ***S-PLMC convergence.*** *Let μ_K^{PLMC} be the distribution of the random variable X_K obtained by iterating **PLMC** Eq. 3, and $\mu_K^{\text{S-PLMC}}$ be the distribution of the random variable Y_K obtained by iteration **S-PLMC** given in Eq. 7. We have TV($\mu_K^{\text{S-PLMC}}, \mu_K^{\text{PLMC}}$) is upper-bounded by:*

$$\sqrt{\frac{\eta}{\lambda}} \left(\sum_{k=0}^{K-1} \mathbb{E}\left\|G(Y_k) - \nabla_x U(Y_k)\right\|^2 + K\beta^2 R^2 \right)^{\frac{1}{2}}.$$

2. ***S-ProxLMC convergence.*** *Let μ_K^{ProxLMC} be the distribution of the random variable X_K obtained by iterating **ProxLMC** Eq. 4, and $\mu_K^{\text{S-ProxLMC}}$ be the distribution of the random variable Y_K obtained by iterating **S-ProxLMC** given in Eq. 8. We have TV($\mu_K^{\text{S-ProxLMC}}, \mu_K^{\text{ProxLMC}}$) is upper-bounded by:*

$$\sqrt{\frac{\eta}{2\lambda}} \left(\sum_{k=0}^{K-1} \mathbb{E}\left\|G(X_k) - \nabla_x U(X_k)\right\|^2 \right)^{\frac{1}{2}}.$$

From Thm. 1, we see that it suffices to approximate the potential gradient $\nabla_x U(X)$ (and not only the potential $U(X)$) in order to guarantee convergence of surrogate LMC. Combining Theorem 1 and bounds in Eqs. 5 and 6 we obtain:

Theorem 2 (Convergence of Surrogate Constrained LMC)**.** *Under assumptions A, B and C we have:*

1. *Assume in S-PLMC that there exists $\delta > 0$ such that $\mathbb{E}\left\|G(Y_k) - \nabla_x U(Y_k)\right\|^2 \leq \delta, \forall k \geq 0$. Set $\lambda = 1$, and $\eta = \tilde{\Theta}(\min(R^2/K, \alpha/K^2))$ where $\alpha = 1/(\delta + \beta^2 R^2)$. Then for $K = \tilde{\Omega}(\varepsilon^{-12} d^{12})$, we have:*
$$TV(\mu_K^{\text{S-PLMC}}, \pi) \leq \varepsilon.$$

2. *Assume in S-ProxLMC that there exists $\delta > 0$ such that $\mathbb{E}\left\|G(X_k) - \nabla_x U(X_k)\right\|^2 \leq \delta, \forall k \geq 0$. Set $\lambda = 1$, and $\eta = \min(\gamma(1 + \beta^2\gamma^2)^{-1}, \frac{1}{\delta K^2})$. Then for $K = \tilde{\Omega}(\varepsilon^{-6} d^5)$ we have:*
$$TV(\mu_K^{\text{S-ProxLMC}}, \pi) \leq \varepsilon.$$

We defer the proofs of the Theorems in Appendix F.

5 Learning Surrogate Gradients

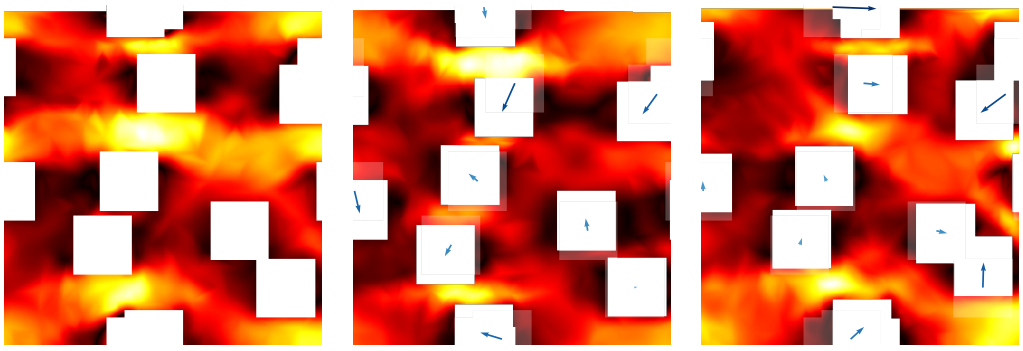
Next, we present two approaches to approximate $\nabla_x U(x)$ with $\mathbf{G}(x)$ using (i) zero-order approximation and (ii) neural network-based Taylor learning (Mukherjee & Zhou, 2006; Mukherjee & Wu, 2006; Wu et al., 2010). We then use either Eq. 7 or Eq. 8 to perform Langevin dynamics. In what follows, we refer to surrogate constrained LMC, as **x-PLMC** or **x-ProxLMC** where **x** is one of four prefixes ({Zero-order, Taylor-2, Taylor-1, Taylor-Reg}).

1. **Zero-order** approximation via Gaussian smoothing (Nesterov & Spokoiny, 2017; Duchi et al., 2015; Ghadimi & Lan, 2013; Shen et al., 2019):
$$G(x) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{U(x + \nu g_j) - U(x)}{\nu}\right) g_j,$$
where $g_1, \ldots g_n$ are i.i.d. standard normal vectors.
2. **Taylor-2**: Given a training set $S = \{(x_i, y_i = \psi(x_i)), x \sim \rho_\Omega, i = 1 \ldots N\}$, we learn two neural networks for the surrogate potential f_θ and gradient G_Λ as in Mukherjee & Zhou (2006):
$$\min_{\theta,\Lambda} \frac{1}{N^2}\sum_{i,j} w_{ij}^\sigma (y_i - f_\theta(x_j) + \langle G_\Lambda(x_i), x_j - x_i\rangle)^2),$$
where $w_{ij}^\sigma = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right)$.
3. **Taylor-1** simplifies the objective function of **Taylor-2** by using only one network f_θ and learn the following:
$$\min_{\theta} \frac{1}{N^2}\sum_{i,j} w_{ij}^\sigma (y_i - f_\theta(x_j) + \langle \nabla_x f_\theta(x_i), x_j - x_i\rangle)^2,$$
4. **Taylor-Reg** uses the Taylor-expansion regularization:
$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 + \lambda R(f_\theta), \text{ where}$$
$$R(f_\theta) = \frac{1}{N^2}\sum_{i,j} w_{ij}^\sigma (y_i - y_j + \langle \nabla_x f_\theta(x_i), x_j - x_i\rangle)^2.$$

The advantage of the Taylor learning of gradients on zero-order estimation is its efficiency at sampling time. Under several mild assumptions, we can show that $\mathbb{E}\left\|G(X_k) - \nabla_x U(X_k)\right\|^2 \leq \delta, \forall k \geq 0$ for each G above.

	Model	Min κ / Respective σ	Min κ / Resp. **satisfied** σ	Per-sam. time (s)
Baseline	Bayesian Opt	0.0552 / 0.5933	0.0846 / 0.4655	1614
Ours	Taylor-Reg-PLMC	0.0613 / 1.4037	0.0732 / 0.4401	952
	Taylor-1-PLMC	0.0544 / 0.8004	0.0963 / 0.4677	852
	Zero-order-PLMC	0.0471 / 0.5594	0.0697 / 0.4764	**15677**
	Taylor-Reg-ProxLMC	0.0639 / 0.8789	**0.0666 / 0.4467**	**856**
	Taylor-1-ProxLMC	0.0548 / 0.6549	0.0876 / 0.4481	972
	Zero-order-ProxLMC	0.0354 / 0.6471	0.0808 / 0.4991	**15080**

Table 1: *Result summary over 20 new samples obtained by our sampling methods on $\pi(x)$ with κ and σ constraints Eq. 6 and the BO baseline. The starting samples are reused from the single constraint case (min $\kappa = 0.0759$, mean $\kappa = 0.1268$, and mean $\sigma = 0.8181$.)*



$\kappa = 0.0871$
$\sigma = 0.4826$

$\kappa = 0.0802$,
$\sigma = 0.6681$

$\kappa = 0.0732$
$\sigma = 0.4401$

Figure 1: *Example of nano-porous structures with corresponding heat flux shown using a color gradient. Yellow regions indicate high phonons flux. The thermal conductivity κ and von Mises stress σ are reported below each structure. The arrows show the moving directions of the pores from their positions on the left structure. (Left) A random sample. (Middle) The sample obtained by Taylor-Reg PMLC starting from the left structure with κ constraint. (Right) The sample obtained by Taylor-Reg PMLC with both κ and σ constraints.*

6 Experiments: Nano-porous Design

We demonstrate the usability of our proposed black-box Constrained Langevin sampling in nano-configuration design under multiple constraints. To design better nano-configurations, we take into account both thermal conductivity and mechanical stability. These physical constraints are respectively specified by thermal conductivity $\kappa(x)$ and mechanical von Mises stress $\sigma(x)$, and they can be obtained by solving the non-linear PDEs (Boltzmann Transport Equation and the continuum linear elasticity Equation respectively). We aim at producing a series of samples x that minimize $\kappa(x)$ to achieve high thermoelectric efficiency while maintaining $\sigma(x)$ lower than some threshold. Based on the posterior regularization formulation in Section 2, we pose the constraint satisfaction as sampling from the following Gibbs distribution:

$$\pi(x) = p_0(x)\frac{\exp(-\lambda_1\kappa(x)^2 - \lambda_2[\sigma(x) - \tau]_+)}{Z}\mathbb{1}_{x\in[0,1]^{20}},$$

where $p_0(x)$ is the uniform distribution over the unit square, which is equivalent to the Poisson process of 10 pores on the square, and τ is a threshold on the maximum value of σ. Sampling from this black box Gibbs distribution is challenging, so our first task is to have good surrogates for the gradient of its potential.

Data. We generate a dataset of 50K nano-porous structures, each of size 100nm $\times$ 100nm. Number of pores is fixed to 10 in this study and each pore is a square with a side length of 17.32nm. We sample the pore centers uniformly over the unit square and construct the corresponding structure after

re-scaling them appropriately. Then, using the solvers OpenBTE (Romano & Grossman, 2015) and Summit ($\sum$MIT Development Group, 2018), we obtain for each structure x a pair of values: thermal conductivity κ and von Mises stress σ. We collect two datasets: $\{(x_i, \kappa_i)\}_{i=1}^N$ and $\{(x_i, \sigma_i)\}_{i=1}^N$, for $N = 50K$ samples.

Features. The pore locations are the natural input features to the surrogate models. Apart from the coordinates, we derive other features based on physical intuitions. We also add pore-pore distances along each coordinate axis as features.

Surrogate gradient methods. We use feed-forward neural networks to model the surrogate gradients, since obtaining their gradients is efficient, thanks to automatic differentiation. We use networks comprised of 4 hidden layers (with ReLu) with sizes $128, 72, 64, 32$ and apply the same architecture to approximate the gradients for κ and σ separately. The output layer is sigmoid. For the Taylor-2 variant , we have an additional output vector. The networks are trained using Adam optimizer with learning rate 10^{-4} and decay 1.0. We fine-tune and select networks with grid-search .

Bayesian optimization as baseline. We use BOTorch library Balandat et al. (2019). The function we wish to optimize is slightly different from the above potential $E(x)$:

$$g(x) = -\kappa(x) - 0.1 \cdot [\sigma(x) - \tau]_+ \quad \text{s.t. } x \in [0, 1]^{20}.$$

We optimize $g(x)$ with BOTorch using QExpected Improvement (qEI) as the acquisition function. We initialize using the same 20 random samples used by our Langevin sampling approach and return 20 new candidates each round. For optimizing the acquisition function, the number of restarts is 20 with 200 samples. The number of samples to estimate the qEI function is 2000. We run BO with 10 steps and report the best result.

Comparison metrics. Starting from 20 samples initialized from $p_0(x)$, we run our proposed black-box Langevin MCs and BO to obtain 20 new realizations from the target distribution $\pi(x)$. To compare outcomes, we report the minimum value of κ and the corresponding σ. We also report the minimum achieved κ when its corresponding σ is below τ.

Discussion. Results are summarized in Table 1. Note that all the surrogate Langevin MCs are initialized from the same set of 20 samples as in BO. In this experiment, we set $\tau = 0.5$, $\lambda_1 = 100$, $\lambda_2 = 1$, the step size $\eta = 1\text{e}{-3}$ and the exponential decay rate 0.8. Our approach can effectively sample new configurations under multiple competing constraints at a significantly low computational cost. Taylor-Reg-ProxLMC achieves the best performance while offering $\mathbf{15}$x speedup over Zero-order PLMC. Compared with BO, our approach achieves higher thermoelectric efficiency. The running time of BO includes data generation and surrogate fitting; however, our approach only requires to fit the surrogate once for all and hence has higher reusability.

REFERENCES

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. Botorch: Programmable bayesian optimization in pytorch. *arXiv preprint arXiv:1910.06403*, 2019.

Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Conference on Learning Theory*, pp. 319–342, 2017.

Bubeck, S., Eldan, R., and Lehec, J. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pp. 1243–1251, 2015.

Chen, G. *Nanoscale energy transport and conversion: a parallel treatment of electrons, molecules, phonons, and photons*.

Chen, M.-H. and Schmeiser, B. Toward black-box sampling: A random-direction interior-point markov chain approach. *Journal of Computational and Graphical Statistics*, 7(1):1–22, 1998.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Durmus, A., Moulines, E., et al. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Ganchev, K., Gillenwater, J., Taskar, B., et al. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049, 2010.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Hu, Z., Yang, Z., Salakhutdinov, R. R., Qin, L., Liang, X., Dong, H., and Xing, E. P. Deep generative models with learnable knowledge constraints. In *Advances in Neural Information Processing Systems*, pp. 10501–10512, 2018.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Lipster, R. and Shiryaev, A. *Statistics of random processes*. Springer, 2001.

Mockus, J. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

Mukherjee, S. and Wu, Q. Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res.*, 2006.

Mukherjee, S. and Zhou, D.-X. Learning coordinate covariances via gradients. *J. Mach. Learn. Res.*, 2006.

Neal, R. M. Slice sampling. *The Annals of Statistics*, 31, 2003.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 2017.

Romano, G. and Di Carlo, A. Multiscale electrothermal modeling of nanostructured devices. *IEEE Trans. Nanotechnol.*, 10(6):1285–1292, 2011. URL `http://ieeexplore.ieee.org/document/5740609/?arnumber=5740609&tag=1`.

Romano, G. and Grossman, J. C. Heat conduction in nanostructured materials predicted by phonon bulk mean free path distribution. *J. Heat Transf.*, 137(7): 071302, 2015. URL `https://heattransfer.asmedigitalcollection.asme.org/article.aspx?articleid=2119334`.

Shen, L., Balasubramanian, K., and Ghadimi, S. Non-asymptotic results for langevin monte carlo: Coordinate-wise and black-box sampling. *arXiv preprint arXiv:1902.01373*, 2019.

$\sum$MIT Development Group, T. $\sum$mit, a scalable computational framework for large-scale simulation of complex mechanical response of materials, 2018. URL `http://summit.mit.edu`.

Wu, Q., Guinney, J., Maggioni, M., and Mukherjee, S. Learning gradients: Predictive models that infer geometry and statistical dependence. *J. Mach. Learn. Res.*, 2010.

A SUPPLEMENTAL EXPERIMENTAL RESULTS

Surrogate gradient methods. We use feed-forward neural networks to model the surrogates because obtaining gradients for such networks is efficient thanks to automatic differentiation frameworks. We use networks comprised of 4 hidden layers with sizes $128, 72, 64, 32$ and apply the same architecture to approximate the gradients for κ and σ separately. The hidden layers compute ReLU activation whereas sigmoid was used at the output layer (after the target output is properly normalized). For the Taylor-2 variant (in Eq. 2), we have an output vector for the gradient prediction. The networks are trained on the corresponding objective functions set up earlier by Adam optimizer with learning rate 10^{-4} and decay 1.0. We fine-tune the networks with simple grid-search and select the best models for comparison.

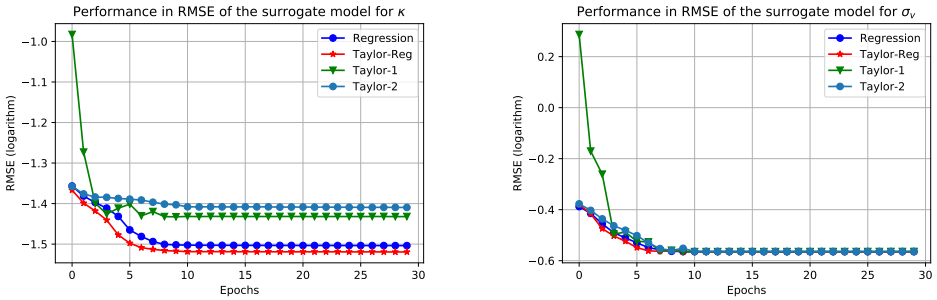


Figure 2: *Comparison of the surrogate variants in testing RMSE. (Left) prediction accuracy for the thermal conductivity κ. (Right) prediction accuracy for mechanical stability σ. Note the difference in scale of κ and σ.*

As emphasized throughout, our focus is more on approximating the gradient rather than learning the true function. However, we need to somehow evaluate the surrogate models on how well they generalize on a hold-out test set. Like canonical regression problems, we compare the surrogate variants against each other using root mean square error (RMSE) on the test set. Figures 2 and 3 shows the results. The left figure shows RMSE for predicting κ and the right one shows RMSE for the von Mises stress σ. We can see that the Taylor-Reg generalizes better and also converges faster than Taylor-1 and Taylor-2 to target RMSE for κ, while all methods result similarly for σ prediction. This is reasonable because the objectives of Taylor-1 and Taylor-2 are not to optimize the mean square error, which we evaluate on here. Figure 3 shows the learning in terms of sample complexity. Again, Taylor-Reg outperforms Taylor-1 and Taylor-2 for κ prediction. In contrast, most models work similarly for σ regression, particularly when the training size is reduced to 50% (25K).

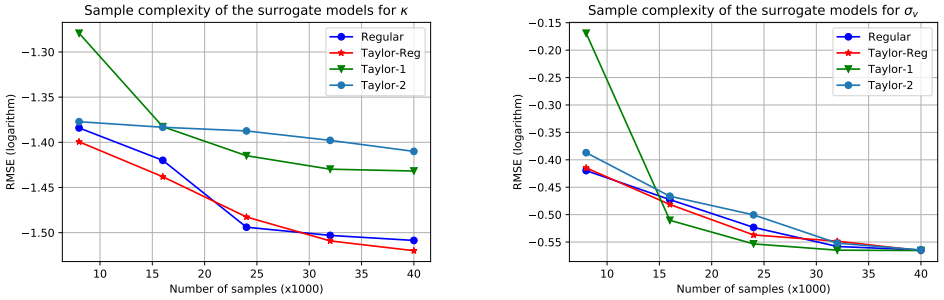
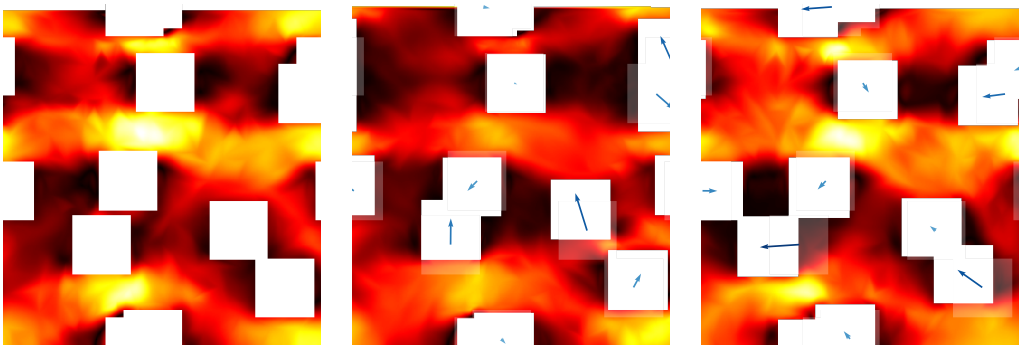


Figure 3: *Comparison of the surrogate models in RMSE on the same test set when the training size is varied. Note the scale difference in the figures due to the different range of values.*

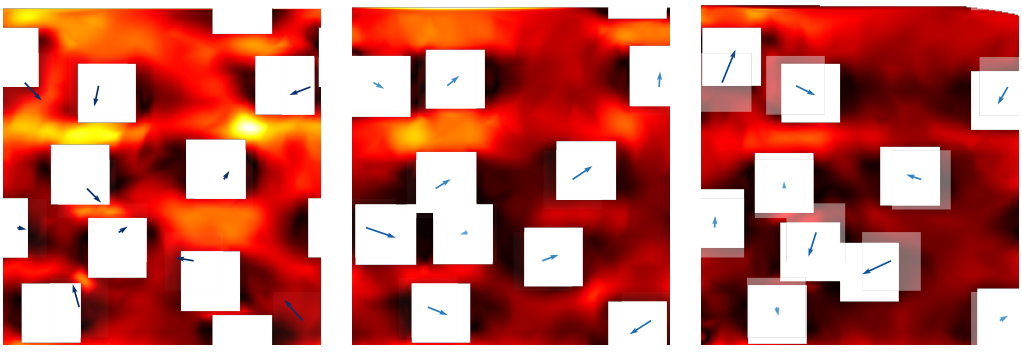
Additional generated samples. We show additional configurations generated by our sampling approach (Taylor-Reg ProxLMC, Taylor-1 ProxLMC and Zero-order ProxLMC) in Fig. 4.

Examples of the samples generated by Zero-order PLMC, Taylor-1 PLMC and the hybrid method are also depicted in Figure 5.



$\kappa = 0.0871, \sigma = 0.4826$ $\kappa = 0.0834, \sigma = 0.4638$ $\kappa = 0.0942, \sigma = 0.4566$

Figure 4: *Example of nano-porous structures with corresponding heat flux shown using a color gradient. Yellow regions indicate high phonons flux. The thermal conductivity κ and von Mises stress σ are reported below each structure. The arrows show the moving directions of the pores. (Left) A random sample. (Middle) The sample obtained by Taylor-Reg ProxLMC starting from the left structure with κ constraint. (Right) The sample obtained by Taylor-Reg ProxLMC with both κ and σ constraints.*



$\kappa = 0.063$ $\kappa = 0.090$ $\kappa = 0.076$

Figure 5: *Samples from by Zero-order PLMC (left), Taylor-1 PLMC (middle) and the hybrid algorithm of Zero-order and Taylor-1 PLMC (right). All are run with the κ constraint.*

B Background on Modeling Nanoscale Heat Transport

At the nanoscale, heat transport may exhibit strong ballistic behaviour and a non-diffusive model must be used (Chen). In this work we use the Boltzmann transport equation under the relaxation time approximation and in the mean-free-path (MFP) formulation (Romano & Grossman, 2015)

$$\Lambda \hat{\mathbf{s}} \cdot \nabla T(\Lambda) + T(\Lambda) = \int \alpha(\Lambda') \langle T(\Lambda') \rangle d\Lambda', \tag{9}$$

where $T(\Lambda)$ is the effective temperature associated to phonons with MFP Λ and direction $\hat{\mathbf{s}}$; the notation $\langle . \rangle$ stands for an angular average. The coefficients $\alpha(\Lambda')$ are given by

$$\alpha(\Lambda') = \frac{K(\Lambda')}{\Lambda'} \left[\int \frac{K(\Lambda'')}{\Lambda''} d\Lambda'' \right]^{-1}, \tag{10}$$

where $K(\Lambda')$ is the bulk MFP distribution. In general, such a quantity can span several orders of magnitude; however, for simplicity we assume the *gray* model, i.e. all phonons travel with the same

MFP, Λ_0. Within this approximation, we have $K(\Lambda) = \kappa_{\text{bulk}}\delta(\Lambda - \Lambda_0)$. In this work we choose Λ_0 = 10 nm, namely as large as the unit cell, so that significant phonons size effects occur. With no loss of generality, we set $\kappa_{\text{bulk}} = 1$ $\text{Wm}^{-1}\text{K}^{-1}$. Eq. 9 is an integro-differential PDE, which is solved iteratively for each phonon direction over an unstructured mesh (Romano & Di Carlo, 2011). We apply periodic boundary conditions along the unit cell while imposing a difference of temperature of ΔT = 1 K along the *x*-axis. At the pores' walls we apply diffusive boundary conditions. Upon convergence, the effective thermal conductivity is computed using Fourier's law, i.e.

$$\kappa_{\text{eff}} = -\frac{L}{\Delta T A}\int_A \mathbf{J}\cdot\hat{\mathbf{n}}dS, \tag{11}$$

where $\mathbf{J} = (\kappa_{\text{bulk}}/\Lambda_0)\langle T(\Lambda_0)\hat{\mathbf{s}}\rangle\hat{\mathbf{n}}$ is the heat flux, L is the size of the unit cell, A is the area of the cold contact (with normal $\hat{\mathbf{n}}$). Throughout the text we use the quantity $\kappa = \kappa_{\text{eff}}/\kappa_{\text{bulk}}$ as a measure of phonon size effects.

C Background on Modeling Mechanical Stress

We model mechanical stress by using the continuum linear elasticity equations

$$\frac{\partial}{\partial x_j}\sigma_{ij} = f_i, \tag{12}$$

where f_i is the body force (which is zero in this case), and σ_{ij} is the stress tensor. Note that we used the Einstein notation, i.e. repeated indexes are summed over. The strain ϵ_{kl} is related to the stress via the fourth-rank tensor elastic constant C_{ijkl}

$$\sigma_{ij} = C_{ijkl}\epsilon_{kl}. \tag{13}$$

The strain is then related to the displacement $\mathbf{u}$ via

$$\epsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial u_k}\right). \tag{14}$$

We apply periodic boundary conditions along the unit-cell and applied solicitation is a small in-plane expansion. Once the stress tensor is calculated, we compute the von Mises stress as

$$\sigma_{VM} = \sqrt{\frac{1}{2}\left(\sigma_3 - \sigma_2\right)^2 + \left(\sigma_3 - \sigma_1\right)^2 + \left(\sigma_2 - \sigma_1\right)^2}, \tag{15}$$

where σ_i are the principal stress axis. As a mechanical stability estimator we use $\sigma = \max_{\mathbf{x}\in D}(\sigma_{VM})$ where D is the simulation domain. To avoid material's plasticity, σ needs to be smaller than the yield stress of a given material. For mechanical simulation we used the SUMIT code ($\sum$MIT Development Group, 2018).

D Background on Stochastic Differential Equations (SDE): Change of Measure and Grisanov's Formula

Theorem 3 (Grisanov Theorem, Change of Measure for Brownian Motion (Lipster & Shiryaev, 2001), Theorem 6.3 page 257)**.** *Let $(W_t, \mathscr{F}_t)$ be a Wiener process (Brownian motion) and $(\beta_t, \mathscr{F}_t)$ a random process such that for any $T > 0$*

$$\int_0^T \|\beta_t\|^2 dt < \infty \text{ a.s}$$

Then the random process : $d\tilde{W}_t = dW_t - \beta_t dt$ or written equivalently: $\tilde{W}_t = W_t - \int_0^t \beta_s ds$, is a Wiener process with respect to $\mathscr{F}_t$, $t \in [0, T]$. Let $P_T^W = \mathscr{L}(W_{[0,T]})$, and $P_T^{\tilde{W}} = \mathscr{L}(\tilde{W}_{[0,T]})$ the densities are given by: $\frac{dP_T^{\tilde{W}}}{dP_T^W} = \exp\left(\int_0^T \langle\beta_s, dW_s\rangle - \frac{1}{2}\int_0^T \|\beta_s\|^2 ds\right)$. It follows that:

$$KL(P_T^W, P_T^{\tilde{W}}) = \frac{1}{2}\mathbb{E}_{P_T^W}\left[\int_0^T \|\beta_s\|^2 ds\right] \tag{16}$$

Theorem 4 (Grisanov Theorem, Change of Measure for Diffusion Processes, (Lipster & Shiryaev, 2001), ()). *Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$*

$$dX_t = \alpha_t(X)dt + dW_t$$
$$dY_t = \beta_t(Y)dt + dW_t$$

where $X_0 = Y_0$ is an $\mathscr{F}_0$ measurable random variable. Suppose that the non-anticipative functionals $\alpha_t(x)$ and $\beta_t(x)$ are such that a unique continuous strong solutions exits for both processes. If for any $T > 0$:

$$\int_0^T \|\alpha_s(X)\|^2 + \|\beta_s(X)\|^2 \, ds < \infty (a.s) \text{ and } \int_0^T \|\alpha_s(Y)\|^2 + \|\beta_s(Y)\|^2 \, ds < \infty (a.s).$$

Let $P_T^X = \mathscr{L}(X_{[0,T]})$, and $P_T^Y = \mathscr{L}(Y_{[0,T]})$.

$$\frac{dP_T^Y}{dP_T^X}(X) = \exp\left(-\int_0^T \langle \alpha_s(X) - \beta_s(X), dX_s \rangle + \frac{1}{2}\int_0^T (\|\alpha_s(X)\| - \|\beta_s(X)\|^2) ds \right).$$

$$KL(P_T^X, P_T^Y) = \frac{1}{2}\mathbb{E}_{P_T^X}\left[\int_0^T \|\alpha_s(X) - \beta_s(X)\|^2 \, ds \right]. \tag{17}$$

E Background on Zero-Order Optimization (Gradient-Free)

Consider the smoothed potential U_ν defined as follows:

$$U_\nu(x) = \mathbb{E}_{g\sim\mathcal{N}(0,I_d)} U(x + \nu g)$$

its gradient is given by:

$$\nabla_x U_\nu(x) = \mathbb{E}_g \frac{U(x+\nu g) - U(x)}{\nu} g,$$

A monte carlo estimate of $\nabla_x U_\nu(x)$ is:

$$\hat{G}_n(x) = \frac{1}{n}\sum_{j=1}^n \left(\frac{U(x+\nu g_j) - U(x)}{\nu} \right) g_j,$$

where $g_1, \dots g_n$ are iid standard Gaussians vectors.

Using known results in zero order optimization under assumptions on smoothness and bounded gradients of the gradients we have for all x ((Nesterov & Spokoiny, 2017; Shen et al., 2019)):

$$\mathbb{E}_g \left\| \hat{G}_1(x) - \nabla_x U(x) \right\|^2 \leq \left(\beta\nu(d+2)^{3/2} + (d+1)^{\frac{1}{2}} \|\nabla_x U(x)\| \right)^2 \leq \left(\beta\nu(d+2)^{3/2} + (d+1)^{\frac{1}{2}} L \right)^2$$

Finally by independence of $u_1, \dots u_n$ we have:

$$\mathbb{E}_{g_1,\dots,g_n} \left\| \hat{G}_n(x) - \nabla_x U(x) \right\|^2 \leq \frac{\left(\beta\nu(d+2)^{3/2} + (d+1)^{\frac{1}{2}} L \right)^2}{n} \tag{18}$$

F Proofs

Proof of Lemma 1. Define the Lagrangian:

$$L(q,\eta) = \int_\Omega \log\left(\frac{q(x)}{p_0(x)} \right) q(x)dx + \sum_{j=1}^{C_e} \lambda_j \int_\Omega (\psi_j(x) - y_j)^2 q(x)dx$$
$$+ \sum_{k=1}^{C_i} \lambda_k \int_{x\in\Omega} (\phi_k(x) - b_k)_+ q(x)dx + \eta\left(1 - \int_{x\in\Omega} q(x) \right)$$

Setting first order optimality conditions on q, we have for $x \in \Omega$:

$$\log\left(\frac{q(x)}{p_0(x)}\right) + 1 + \sum_{j=1}^{C} \lambda_j(\psi_j(x) - y_j)^2 + \sum_{k=1}^{C_i} \lambda_k(\phi_k(x) - b_k)_+ - \eta = 0$$

Hence we have:

$$q(x) = p_0(x)\frac{\exp\left(-\sum_{j=1}^{C_e} \lambda_j(\psi_j(x) - y_j)^2 - \sum_{k=1}^{C_i} \lambda_k(\phi_k(x) - b_k)_+\right)}{e\exp -\eta}, x \in \Omega$$

and

$$q(x) = 0, x \notin \Omega,$$

First order optimality on η give us: $\int_\Omega q(x) = 1$, we conclude by setting $e\exp(-\eta) = Z$. □

Proof of Theorem 1 1) Projected Langevin. Let us define the following continuous processes by interpolation of X_k and Y_K (Piecewise constant):

$$d\tilde{X}_t = P_\Omega(\tilde{U}_t(\tilde{X})dt + \sqrt{2\lambda}dW_t)$$

where $\tilde{U}_t(\tilde{X}) = -\sum_{k=0}^{\infty} \nabla_x U(\tilde{X}_{k\eta})\mathbb{1}_{t\in[k\eta,(k+1)\eta]}(t)$. Similarly let us define :

$$d\tilde{Y}_t = P_\Omega(G_t(\tilde{Y})dt + \sqrt{2\lambda}dW_t)$$

where $G_t(\tilde{Y}) = -\sum_{k=0}^{\infty} G(\tilde{Y}_{k\eta})\mathbb{1}_{t\in[k\eta,(k+1)\eta]}(t)$.

It is easy to see that we have : $X_k = \tilde{X}_{k\eta}$ and $Y_k = \tilde{Y}_{k\eta}$.

Let $\pi^T_{\tilde{X}}$ and $\pi^T_{\tilde{Y}}$ be the distributions of $(\tilde{X}_t)_{t\in[0,T]}$ and $(\tilde{Y})_{t\in[0,T]}$.

Note that :

$$d\tilde{Y}_t = P_\Omega\left(\tilde{U}_t(\tilde{X}_t)dt + \sqrt{2\lambda}(dW_t + \frac{1}{\sqrt{2\lambda}}(G_t(\tilde{Y}_t) - \tilde{U}_t(\tilde{X}_t))dt)\right)$$

Let

$$d\tilde{W}_t = dW_t + \frac{1}{\sqrt{2\lambda}}(G_t(\tilde{Y}_t) - \tilde{U}_t(\tilde{X}_t))dt$$

Hence we have :

$$d\tilde{Y}_t = P_\Omega\left(\tilde{U}_t(\tilde{X}) + \sqrt{2\lambda}d\tilde{W}_t\right),$$

Assume that $X_0 = Y_0$ there exists $\mathcal{Q}$ such that , $X_T = \mathcal{Q}(\{W_t\}_{t\in[0,T]})$ and $Y_T = \mathcal{Q}((\tilde{W}_t)_{t\in[0,T]})$. Let $\mu^{\tilde{X}}_T$ be the law of $\tilde{X}_{t\in[0,T]}$. Same for $\mu^{\tilde{Y}}_T$. The proof here is similar to the proof of Lemma 8 in (Bubeck et al., 2015). By the data processing inequality we have:

$$\text{KL}(\mu^{\tilde{X}}_T, \mu^{\tilde{Y}}_T) \leq \text{KL}(W_{t\in[0,T]}, \tilde{W}_{t\in[0,T]}),$$

Now using Grisanov's Theorem for change of measure of Brownian Motion (Theorem 3) we have:

$$\text{KL}(W_{t\in[0,T]}, \tilde{W}_{t\in[0,T]}) = \frac{1}{4\lambda}\mathbb{E}\int_0^T |G_t(\tilde{Y}_t) - \tilde{U}_t(\tilde{X}_t)|^2 dt$$

Consider $T = K\eta$, hence we have (with some abuse of notation we drop tilde as $Y_k = \tilde{Y}_{k\eta}$):

$$\begin{aligned}
\mathrm{KL}(\mu_T^{\tilde{X}}, \mu_T^{\tilde{Y}}) &\leq \frac{1}{4\lambda}\mathbb{E}\int_0^{K\eta} |G_t(\tilde{Y}_t) - \tilde{U}_t(\tilde{X}_t)|^2 dt \\
&= \frac{1}{4\lambda}\mathbb{E}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \left\|G(Y_{k\eta}) - \nabla_x U(X_{k\eta})\right\|^2 dt \\
&= \frac{\eta}{4\lambda}\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(X_{k\eta})\right\|^2 \\
&= \frac{\eta}{4\lambda}\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(Y_{k\eta}) + \nabla_x U(Y_{k\eta}) - \nabla_x U(X_{k\eta})\right\|^2 \\
&\leq \frac{\eta}{2\lambda}\sum_{k=0}^{K-1}\left(\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(Y_{k\eta})\right\|^2 + \mathbb{E}\left\|\nabla_x U(Y_{k\eta}) - \nabla_x U(X_{k\eta})\right\|^2\right)
\end{aligned}$$

where in the last inequality we used the fact that $||a-b||^2 \leq 2(||a||^2 + ||b||^2)$. Note that we have by smoothness assumption on U:

$$\left\|\nabla_x U(Y_{k\eta}) - \nabla_x U(X_{kh})\right\|^2 \leq \beta^2 \left\|X_{kh} - Y_{kh}\right\|^2$$

Let R be the diameter of Ω, we can get a bound as follows:

$$\begin{aligned}
\mathrm{KL}(\mu_T^{\tilde{X}}, \mu_T^{\tilde{Y}}) &\leq \frac{\eta}{2\lambda}\left(\underbrace{\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(Y_{k\eta})\right\|^2}_{\text{Gradient approximation error}} + \beta^2 \sum_{k=0}^{K-1}\mathbb{E}\left\|X_{kh} - Y_{kh}\right\|^2\right) \\
&\leq \frac{\eta}{2\lambda}\left(\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(Y_{k\eta})\right\|^2 + K\beta^2 R^2\right)
\end{aligned}$$

Now using Pinsker inequality we have:

$$TV(\mu_T^{\tilde{X}}, \mu_T^{\tilde{Y}})^2 \leq 2\mathrm{KL}(\mu_T^{\tilde{X}}, \mu_T^{\tilde{Y}}) \leq \frac{\eta}{\lambda}\left(\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_{k\eta}) - \nabla_x U(Y_{k\eta})\right\|^2 + K\beta^2 R^2\right)$$

Hence for $T = K\eta$ we have:

$$TV(\mu_K^{\text{S-PLMC}}, \mu_K^{\text{PLMC}}) \leq \sqrt{\frac{\eta}{\lambda}}\left(\sum_{k=0}^{K-1}\mathbb{E}\left\|G(Y_k) - \nabla_x U(Y_k)\right\|^2 + K\beta^2 R^2\right)^{\frac{1}{2}}. \tag{19}$$

□

Proof of Theorem 1 2) Proximal LMC. Let us define the following continuous processes by interpolation of X_k and Y_K (Piecewise constant):

$$d\tilde{X}_t = \tilde{U}_t(\tilde{X})dt + \sqrt{2\lambda}dW_t$$

where $\tilde{U}_t(\tilde{X}) = -\sum_{k=0}^{\infty}(\nabla_x U(\tilde{X}_{k\eta}) + \frac{1}{\gamma}(\tilde{X}_{k\eta} - P_\Omega(\tilde{X}_{k\eta})))\mathbb{1}_{t\in[k\eta,(k+1)\eta]}(t)$. Similarly let us define :

$$d\tilde{Y}_t = G_t(\tilde{Y})dt + \sqrt{2\lambda}dW_t$$

where $G_t(\tilde{Y}) = -\sum_{k=0}^{\infty}(G(\tilde{Y}_{k\eta}) + \frac{1}{\gamma}(\tilde{Y}_{k\eta} - P_\Omega(\tilde{Y}_{k\eta})))\mathbb{1}_{t\in[k\eta,(k+1)\eta]}(t)$. Now applying Grisanov's Theorem for diffusions (Theorem 4) we have:

$$\begin{aligned}
\mathrm{KL}(\mu_T^{\tilde{X}}, \mu_T^{\tilde{Y}}) &= \frac{1}{4\lambda}\mathbb{E}_{P_T^X}\left[\int_0^T \left\|U_t(\tilde{X}) - G_t(\tilde{X})\right\|^2 dt\right] \\
&= \frac{1}{4\lambda}\mathbb{E}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\left\|G(\tilde{X}_{k\eta}) - \nabla_x U(\tilde{X}_{k\eta})\right\|^2 dt \\
&= \frac{\eta}{4\lambda}\sum_{k=0}^{K-1}\mathbb{E}\left\|G(\tilde{X}_{k\eta}) - \nabla_x U(\tilde{X}_{k\eta})\right\|^2 \\
&= \frac{\eta}{4\lambda}\sum_{k=0}^{K-1}\mathbb{E}\left\|G(X_k) - \nabla_x U(X_k)\right\|^2.
\end{aligned}$$

Now using Pinsker inequality we have:

$$TV(\mu_{\tilde{X}}^T, \mu_{\tilde{Y}}^T)^2 \leq 2\mathrm{KL}(\mu_{\tilde{X}}^T, \mu_{\tilde{Y}}^T).$$

Hence for $T = K\eta$ we have:

$$TV(\mu_K^{\text{S-ProxLMC}}, \mu_K^{\text{ProxLMC}}) \leq \sqrt{\frac{\eta}{2\lambda}}\left(\sum_{k=0}^{K-1}\mathbb{E}\left\|G(X_k) - \nabla_x U(X_k)\right\|^2\right)^{\frac{1}{2}}. \tag{20}$$

□

Proof of Theorem 2 . **S-PLMC.** If we set $\lambda = 1$, $\eta \leq \alpha/K^2$, where $\alpha = 1/(\delta + \beta^2 R^2)$, in this Corollary we obtain that : $TV(\mu_K^{S-PLMC}, \mu_K^{PLMC}) \leq \frac{1}{\sqrt{K}}$. Assuming A, B and C we consider $\eta \leq \min(R^2/K, \alpha/K^2)$, and $K = \tilde{\Omega}(\varepsilon^{-12}d^{12})$. Now using the triangle inequality together with the bounds in Eq.s 5 we have: $TV(\mu_K^{S-PLMC}, \pi) \leq TV(\mu_K^{S-PLMC}, \mu_K^{PLMC}) + TV(\mu^{PLMC}, \pi) \leq \varepsilon + \frac{1}{\sqrt{K}}$.

S-ProxLMC. We conclude with a similar argument for $TV(\mu_K^{S-ProxLMC}, \pi)$ using Eq.s 6. Considering $\eta = \min(\gamma(1+\beta^2\gamma^2)^{-1}, \frac{1}{\delta K^2})$, and $K = \tilde{\Omega}(\varepsilon^{-6}d^5)$, we obtain $(\varepsilon + \frac{1}{\sqrt{K}})$ approximation in TV of the target Gibbs distribution.

□